Brügelmann, Hans

"Zu Risiken und Nebenwirkungen...". Warnung vor einer naiven "Evidenzbasierung" in der Pädagogik

Lehren und lernen 7 (2019) 5, S. 29-34

Hans Brügelmann

„Zu Risiken und Nebenwirkungen …": Warnung vor einer naiven „Evidenzbasierung" in der Pädagogik[1]

Yong Zhao hat in seinem erhellenden Buch „Was wirkt, kann auch schaden" (2018) das gegenwärtig dominierende Paradigma empirischer Bildungsforschung kritisch unter die Lupe genommen und dessen beide Hauptmethoden: zum einen Bestandsaufnahmen in großen, möglichst repräsentativen Stichproben, zum andern die Evaluation von Interventionen in Unterrichtssettings (im Rahmen von Feldstudien oder quasi-experimentell, jeweils mit Kontrollgruppen). *Zhaos* Befunde bezüglich einer Reihe großer politisch motivierter Programme/Interventionen sind überaus ernüchternd: Wirkungen sind nie einsinnig, sondern in der Regel ambivalent; Kontexte bewirken beträchtliche Effekt-Streuungen und wurden zu wenig berücksichtigt; kurzfristigen Erfolgen stehen längerfristige Nachteile gegenüber; Nebenwirkungen werden vernachlässigt. Gegenüber den hierzulande angesagten Qualitätsinitiativen ist höchste Skepsis angesagt!

▸ Stichwörter: Evaluierung, empirische Bildungsforschung, Wirkungsforschung, Evidenzbasierung

Ende September letzten Jahres hat die brandenburgische Kultusministerin *Britta Ernst* den Einsatz der Methode *Lesen durch Schreiben* für die Schulen ihres Landes verboten. Sie stützte sich dabei auf Vorabveröffentlichungen aus einem empirischen Vergleich dieses Ansatzes mit einer nicht näher spezifizierten Fibel, dessen Fazit gelautet hatte: „Insgesamt kann nach den Ergebnissen dieser längsschnittlichen wie querschnittlichen Analysen ein Rechtschreibunterricht mit … Lesen durch Schreiben … nicht empfohlen werden" (Kuhl/Röhr-Sendlmeier 2018a). Zwar hatten die Autorinnen dieses Urteil nach deutlicher Kritik an der Pauschalität ihrer Folgerung und der fehlenden Überprüfbarkeit ihrer empirischen Grundlagen inzwischen erheblich eingeschränkt („kann … nicht uneingeschränkt empfohlen werden", Kuhl/Röhr-Sendlmeier 2018b), aber das Grundproblem bleibt: Darf eine Ministerin bzw. Schulverwaltung Unterrichtsmethoden verbieten oder vorschreiben? Und bieten empirische Vergleichsuntersuchungen dafür eine tragfähige Grundlage? Welche Aussagekraft, welchen Geltungsbereich haben also die Befunde der seit 2000 sowohl auf System- als auch auf Unterrichtsebene zunehmenden Evaluationsstudien?

Dies sind Kernfragen in *Yong Zhaos* erhellendem Buch „Was wirkt, kann auch schaden" (2018). Es überzeugt als eine fachlich fundierte und differenziert urteilende, aber zugleich anschauliche Auseinandersetzung mit dem gegenwärtig dominierenden Paradigma empirischer Bildungsforschung und dessen beiden Hauptmethoden: zum einen Bestandsaufnahmen in großen, möglichst repräsentativen Stichproben, zum andern die Evaluation von Interventionen in Unterrichtssettings (im Rahmen von Feldstudien oder quasi-experimentell, jeweils mit Kontrollgruppen). Die Erkenntnisse bestehen in statistischen Durchschnittswerten, die mit der Hilfe von standardisierten Tests gewonnen wurden.

Für eine „evidenzbasierte" Praxis, die sich an diesem sog. „Goldstandard" orientiert, wird der Pädagogik immer wieder die Medizin als Vorbild vorgehalten (s. aber die Kritik bei Sackett et al. 2000). Auch *Zhao* wählt viele Bilder und konkrete Beispiele aus diesem Bereich. Sein Ziel ist aber nicht, den Vergleich von statistischen Kennwerten allein zur Grundlage bildungspolitischer oder didaktischer Entscheidungen zu machen. In seinen Analogien zu evidenzbasierter Medizin erkennt *Zhao* das Potenzial dieses Forschungsansatzes durchaus an, konzentriert sich aber auf die Grenzen und auf häufige Schwächen in seiner Umsetzung durch die Forscher/innen und in der Rezeption durch die Politik (weniger durch die Schulpraxis). Sein Ausgangspunkt sind pharmazeutische Katastrophen wie Contergan, nach denen den Entwickler/innen von Medikamenten für deren Zulassung die Dokumentation von Nebenwirkungen zur Auflage gemacht worden ist (S. 33ff.).

Zhaos Hauptkritik: Bisher hat die Bildungsforschung nach den im Durchschnitt wirksamsten Programmen bzw. Methoden gesucht – gemessen an eng definierten Fachzielen. Stattdessen sollte sie untersuchen und berichten, unter welchen Bedingungen und für welche Gruppen einzelne Programme/Methoden besonders ertragreich sind bzw. welche unerwünschten Nebenwirkungen sie jeweils haben – über die von den Entwicklern selbst angestrebten Ziele hinaus. Anders gesagt: Allgemeinurteile wie „gut" oder „schlecht" seien wenig hilfreich. Bildungspolitik und

Schulpraxis brauchen stattdessen Informationen über Potenziale und Risiken von Interventionen, deren Entfaltung von der Umsetzung vor Ort abhängt, und vom Kontext, in dem sie jeweils eingesetzt werden.

Es sind vor allem vier Probleme, die es bei der Beurteilung von pädagogischen Interventionen als mehr oder weniger „gut" gibt:

- Was auf einzelne Personen in einer Zieldimension positiv wirkt, kann ihnen in einer anderen schaden (S. 78ff.).
- Was bestimmten (Gruppen von) Schüler/innen hilft, kann auf andere (Gruppen) negativ wirken. (S. 89ff., 101ff.).
- Was in einigen Klassen, Schulen oder Regionen erfolgreich umgesetzt wird, kann in anderen scheitern (S. 44f., 99ff.).
- Was kurzfristig sehr wirksam erscheint, kann sich langfristig als wenig förderlich erweisen (S. 54, 85 ff.).

Diese differenziellen Wirkungen bildungspolitischer und didaktischer Interventionen diskutiert *Zhao* im Detail an konkreten Schulreformen und prominenten Beispielen aus der Bildungsforschung.

Fehlende Aufmerksamkeit für mögliche Nebenwirkungen: *No Child left behind*

Zhaos erstes Beispiel für eine naive Bildungsreform, die ohne Prüfung möglicher Nebenwirkungen in Gang gesetzt wurde, ist das 2002 in den USA initiierte Programm *„No child left behind"* (NCLB), mit dessen Hilfe einerseits das Leistungsniveau im Lesen und in Mathematik insgesamt angehoben, andererseits die Kluft zwischen leistungsschwachen und leistungsstarken Schülerinnen geschlossen werden sollte (S. 7ff.).

Als Patentlösung für die genannten Probleme wurden folgende Maßnahmen gesetzlich verankert (S. 9ff.):

- an Lernfortschritte in Leistungstests gebundene Rechenschaftspflichten der Schulen *(„high-stakes tests")*
- freie Schulwahl für Schüler/innen bzw. ihre Eltern
- Zulassung nur von empirisch bewährten Methoden und Programmen
- fachliche Qualifikation von Lehrer/innen.

Obwohl einige Bundesstaaten die Anforderungen an die angestrebte *„proficiency"* senkten, um die Ziele leichter zu erreichen (S. 11), wurden diese durchweg verfehlt. So war – gemessen an den parallel laufenden landesweiten NAEP-Leistungs-Tests – 2015 immer noch mehr als die Hälfte der Schüler/innen nicht *„proficient"* und die Abstände zwischen den sozio-ökonomischen und kulturellen Gruppen konnten bei Weitem nicht geschlossen werden (S. 12ff.).

Stattdessen wurden erhebliche negative Nebenwirkungen der Maßnahmen festgestellt (S. 15ff.).:

- Verengung des Curriculums, vor allem in Schulen, die mit benachteiligten Schüler/innen arbeiteten
- verbreitetes *„teaching to the test"* – ohne positive Effekte auf Testleistungen
- Fokussierung des Unterrichts auf Schüler/innen, deren Leistungsentwicklung versprach, den Testdurchschnitt für die Lerngruppe am stärksten anzuheben
- mehr Prüfungsangst bei den *High-stakes*-Tests
- durch Konzentration auf kurzfristigen Testerfolg Verzicht auf bessere Förder-Alternativen.

Zhaos Fazit: „Zusammenfassend lässt sich sagen, dass NCLB eine totale Katastrophe war. Es hat die angestrebten Ziele nicht erreicht, selbst nicht in deren engster Definition: der Verringerung der Lücke in den Testergebnissen beim Lesen und Mathe zwischen benachteiligten Kindern und ihren privilegierten Kameraden. Aber es war äußerst wirksam dabei, unerwünschte Ergebnisse zu erzielen, die sich negativ auf Schüler, Lehrer, Bildungseinrichtungen und Kultur auswirkten. (…) Die Antwort darauf, warum das Gesetz beim Erreichen des angestrebten Ergebnisses nicht wirksam war, ist ziemlich einfach: Seine Diagnose der Ursache der Leistungslücke war falsch. Die Kluft war eher das Ergebnis von Armut, Segregation, Rassismus und mangelnden Investitionen in die Bildung benachteiligter Menschen, anstatt unqualifizierter Erzieher/innen, die als unwillig und unfähig galten, Kinder gut zu unterrichten. Eine falsche Diagnose führte zu einer unwirksamen Behandlung." (S. 21f.)

Durchschnittswerte verdecken Streuung von Effekten: *Reading First*

Ein zentraler Pfeiler von NCLB war die Bundesinitiative *Reading First*, die ab 2002 von den Bundesstaaten und Schulbezirken forderte, im Leseunterricht nur *„wissenschaftlich bewährte"* Programme zu nutzen, und deren Einführung finanziell förderte. Basis des Programms waren die Empfehlungen des *National Reading Panels* von 2000: „Der beste Ansatz für den Unterricht im Lesen ist ein Konzept, das Folgendes beinhaltet: explizite Unterweisung in phonemischer Bewusstheit, systematische Unterweisung in Laut-Buchstaben-Beziehungen, Methoden zur Verbesserung der Leseflüssigkeit und Maßnahmen zur Verbesserung des Textverständnisses." (S. 26) Aber die Evaluation von *Reading First* zeigte bis 2008 keine Lerngewinne (S. 24).

Das lag einerseits an Korruption, da nämlich bestimmte Verlagsprodukte von der Politik bevorzugt wurden, obwohl sie nicht positiv evaluiert waren (S. 26). Aber auch „wissenschaftlich bewährte" Programme, die konzept-

gerecht umgesetzt wurden, wurden nicht in den Schulen so akzeptiert wie erwartet (S. 27f.). Nach *Zhao* liegt das an den sehr eingeschränkten Zielkriterien des National Reading Panels und daran, dass nur Studien ausgewertet wurden, die dem oben erwähnten „Goldstandard" empirischer Bildungsforschung entsprachen (S. 28f.).

> Um Unterschiede im Lernerfolg tatsächlich einer Intervention zurechnen zu können, verlangt der „Goldstandard" experimenteller Forschung, dass Schüler/innen und Lehrer/innen per Zufall einer Versuchs- und einer Kontrollgruppe zugewiesen werden (*randomized controlled trials* = RCTs). Im Schulalltag ist das in der Regel nicht möglich. Aber es wäre auch nur begrenzt aussagekräftig. So wichtig empirische Belege sind, um die Qualität eines Programms oder einer Methode einzuschätzen, RCTs erfassen nur einen Ausschnitt relevanter Informationen: Denn „das Problem mit RCT ist, dass es viel darüber aussagt, ob eine Behandlung wirksam ist, um ein bestimmtes Ergebnis zu erzielen, aber nichts darüber, ob das Ergebnis wünschenswert ist, noch darüber, ob die Maßnahme auch andere Ergebnisse verursacht" (S. 120).

Bei seiner Suche nach möglichen Erklärungen für den Misserfolg von *Reading First* stellt *Zhao* eine Reihe von Hypothesen über mögliche Nebenwirkungen auf, die in den Studien bzw. ihrer Auswertung durch das *National Reading Panel* nicht erfasst wurden:

„Die einzelnen Studien, die vom *National Reading Panel* untersucht wurden, mögen positive Auswirkungen auf einige Leistungen gezeigt haben, aber *Reading First* könnte in anderen Bereichen Schaden zugefügt haben, wie z. B. einen Motivations- und Interessenverlust. Es hat vielleicht bei einigen Schülern funktioniert und könnte anderen geschadet haben. Es kann eine völlige Zeitverschwendung für einige Schüler/innen gewesen sein, die von anderen Konzepten hätten profitieren können, ein potenzieller Schaden, der einer Fehldiagnose einer Krankheit oder einer falschen Anwendung von Medikamenten gleichkommt." (S. 37)

Wie gesagt, diese Annahmen lassen sich anhand der vorliegenden Studien nicht überprüfen – eine verbreitete Schwäche von „evidenzbasierter" Evaluation. Denn damit fehlen wesentliche Daten für eine differenzierte Einschätzung von Stärken und Schwächen der untersuchten Programme oder Methoden.

Schon bei Studien in den Bereichen Landwirtschaft und Kernenergie hatte sich gezeigt, dass eindimensionale Wirkungsanalysen nicht ausreichen, um den nachgewiesenen Nutzen einer Entscheidung mit den Kosten, zu denen eben auch Nebenwirkungen zählen, abzuwägen (S. 29). Kontrollierte Studien können meist nur kurzfristige Effekte vergleichen (S. 39), Stärken bzw. Schwächen nur in wenigen ausgewählten Zielen nachweisen (S. 37), und sie erfassen nicht differenzielle Wirkungen in unterschiedlichen Kontexten oder auf verschiedene Gruppen (S. 30). Diese Kritik ist nicht neu. Sie war schon seit Anfang der 1970er-Jahre in Kontroversen über – für den pädagogischen Bereich – angemessene Evaluationskonzepte diskutiert worden (House 1980, Elliott/Kushner 2007).

Nebenwirkungen und Kosten positiver Effekte: DISTAR und *direct instruction* allgemein

Ende der 1960er-Jahre hatte US-Präsident *Johnson* *„Follow Through"* in Gang gesetzt, ein bundesweites Grundschulprogramm kompensatorischer Förderung im Anschluss an *„Head Start"* im Kindergarten. Im Rahmen dieses Programms wurden didaktisch sehr unterschiedlich angelegte Projekte finanziert, die auch vergleichend evaluiert wurden. Ihr Erfolg wurde mit standardisierten Tests gemessen, die vor allem kognitive Ziele, u. a. Basisfertigkeiten im Lesen und Rechnen, erfassten. Mehrfach erwies sich dabei eine direkte Instruktion (DI) in eben diesen grundlegenden Fertigkeiten als erfolgreich. Allerdings gab es große Unterschiede zwischen verschiedenen Einrichtungen, die mit demselben Programm arbeiteten (S. 44f.). Außerdem wurden von Eltern trotzdem häufig offenere Programme bevorzugt. Der Grund: Programme wie DISTAR (und Methoden der direkten Instruktion generell) wirken nicht mechanisch, ihr Erfolg ist vielmehr situationsabhängig. Und neben den positiven Effekten im unmittelbaren Zielbereich scheinen für die Betroffenen weitere Kriterien wichtig (gewesen) zu sein, die von der Evaluation nicht erfasst wurden.

Dass Erfolge didaktischer Programme ihren Preis haben, war schon in der Auswertung einer 1920 von *Terman* begonnenen Längsschnittstudie sichtbar geworden. Dabei stellten *Margaret Kern* und *Howard S. Friedman* 2009 fest: „Frühes Lesen ging mit frühem Bildungserfolg einher, war aber auch verbunden mit schlechteren langfristigen Ergebnissen, einschließlich niedrigerem allgemeinem Bildungsniveau, schlechterer Anpassung im Jugendlichen- und Erwachsenenalter und erhöhtem Alkoholkonsum." (zitiert S. 86) Auch wenn statistische Zusammenhänge noch nicht Kausalität bedeuten, regen die Ergebnisse dazu an, sich Gedanken über ein breiteres Spektrum von Zielen zu machen und auch die Nachhaltigkeit von Effekten zu bedenken. So berichtet *Zhao* Befunde aus Studien von *Bonawitz* et al., *Buchsbaum* et al., *Dehn/Kuhn* und *Kapur*, die zeigen, dass eine direkte Instruktion bei der raschen Vermittlung von Wissen im angezielten Leistungsbereich durchaus erfolgreich sein kann und gleichzeitig mit ungünstigen Auswirkungen auf exploratives und kreatives Verhalten und – vor allem längerfristig – beim Transfer auf neue Situationen zu rechnen ist (S. 46, 51f., 54).

Zhaos Fazit: „DI kann in der Wissensvermittlung effektiv sein *und gleichzeitig* Kreativität und Neugierde unterdrücken." (S. 54)

Vernachlässigung von Zielkonkurrenzen: PISA und der Erfolg der ostasiatischen Tigerstaaten

Nach dem vorzüglichen Abschneiden ostasiatischer Staaten bei PISA haben eine Reihe westlicher Bildungspolitiker/innen und Pädagog/innen dort ihr neues Mekka entdeckt. Beispielsweise wird in England in der Hälfte der Grundschulen Mathematik nach einer Methode unterrichtet, die in Hong Kong, Singapur und Shanghai praktiziert wird, und zusätzlich wurden Schulbücher für den Mathematikunterricht aus China importiert (S. 57). Aber ist diese Vorbildrolle berechtigt – und lässt sie sich aus der PISA-Studie tatsächlich ableiten?

Zhao sieht die häufig geäußerte Kritik an Design und Methodik der internationalen Leistungsvergleiche zwar grundsätzlich als berechtigt an (S. 59). Aber er verweist nur zusammenfassend auf entsprechende Publikationen (wo man dann z. B. findet, dass die Aufgabentexte auf Deutsch fast 20% länger sind, so dass Schüler/innen in der festgesetzten Zeit gar nicht so viele Aufgaben bearbeiten können wie ihre englischsprachigen Konkurrent/innen in anderen Ländern, womit ein Vergleich der Anzahl richtiger Lösungen wenig aussagekräftig ist). Dennoch lassen sich nach seiner Einschätzung die PISA-Erfolge Ostasiens damit nicht in Frage stellen. Selbst in China aufgewachsen, erklärt er sie mit besonderen kulturellen und rechtlichen Bedingungen in den genannten Ländern (S. 60ff.): Zentral erlassene, detaillierte und durch Tests immer wieder kontrollierte Lehrpläne steuern den Unterricht in strikter Form; zudem fallen die aufgewandte Zeit für den Unterricht und die Hausaufgaben in den Hauptfächern, aber auch die finanziellen Investitionen der Eltern in private Zusatzförderung viel höher aus als in Europa oder Nordamerika – vor allem bedingt durch das in der Gesellschaft tief verankerte konfuzianische Bildungsideal. Zugleich macht *Zhao* aber darauf aufmerksam, dass umgekehrt vielen in Ostasien die westlichen Bildungssysteme bzw. ihre pädagogischen Prinzipien als Vorbild erscheinen (S. 56, 65ff.). Wie ist das möglich?

Bei PISA und vielen anderen Studien zur Qualität von Bildungssystemen oder zu konkreten Konzepten sind die Schülerleistungen in wenigen Fächern und innerhalb dieses engen Kanons nur in ausgewählten, weil leicht testbaren Kompetenzbereichen untersucht worden. Persönlichkeitsentwicklung, Sozialverhalten, Neugier, Kreativität, Selbstvertrauen, Lebensfreude und viele andere Ziele geraten dabei aus dem Blick. In diesen Bereichen gibt es gerade in den ostasiatischen Ländern Probleme, die *Zhao* mit Verweis auf entsprechende Befunde als „Kosten" des PISA-Erfolgs aufrechnet (S. 66ff.). Generell schneiden bei PISA und TIMSS Länder mit hohen Leistungen in den Tests tendenziell schlecht ab, wenn es um die Einstellung zu den entsprechenden Fächern geht (S. 69).

Zhaos Fazit: „Die ostasiatischen Systeme sind in der Tat sehr effektiv, um in einer begrenzten Anzahl von Fächern hervorragende Testergebnisse zu erzielen. Die herausragenden Leistungen bei Tests gehen aber einher mit weniger Selbstvertrauen, weniger Zufriedenheit und weniger Kreativität sowie weniger Vielfalt an Talenten." (S. 74)

Diese eindimensionale Bewertung von Lern„erfolgen" ist ein generelles Problem.

Übersehen von Interaktionseffekten: *Hatties Meta-Meta-Analyse*

Hatties 2008 veröffentlichte Ranglisten „erfolgreicher" Methoden haben die bildungspolitischen und didaktischen Diskussionen der letzten Jahre in vielen Ländern maßgeblich bestimmt. Wie bei PISA benennt *Zhao* auch hier eine Reihe kritischer Punkte im methodischen Vorgehen (S. 76f.), konzentriert sich dann aber auf ein grundsätzlicheres Problem: Die Ranglisten „Erfolg" versprechender Methoden und Bedingungen seien wegen der breiten Palette verschiedener Ziele, deren Verhältnis zueinander zudem sehr unterschiedlich aussehen kann (S. 81), nur wenig aussagekräftig.

Zhao benennt mit Analogien aus der Ökologie vier sehr unterschiedliche Ziel-Konstellationen (S. 79ff.):

- *Konkurrenz* bedeutet, dass die Nutzung der knappen Ressourcen einer Dimension zu Schwächen in einer anderen führt: Bei NCLB ging schon die zeitliche Konzentration auf die getesteten Schwerpunkte zu Lasten von „Nebenfächern" wie Sozialkunde, Literatur oder Musik.
- Als *Plünderung* bezeichnet er die Unverträglichkeit von zwei Zielen: Wer hohe fachliche Leistungen in standardisierten Test favorisiert, nimmt in Kauf, dass die Kreativität der Betroffenen leidet.
- Ein *verträgliches Nebeneinander* liegt vor, wenn eine Stärke von einer anderen profitiert, ohne dass dies auch umgekehrt gilt: Hartnäckigkeit und Ausdauer fördern eine höhere fachliche Kompetenz, aber diese wirkt sich nicht auf die genannten Einstellungen aus.
- Als Beispiel für eine *wechselseitige Stärkung* nennt *Zhao* Selbstbestimmung und Wohlbefinden.

Die ersten beiden Situationen sind charakteristisch für negative Nebenwirkungen, die bei der Beurteilung von „erfolgreichen" Interventionen bedacht werden müssen. Als besonders problematisch sieht *Zhao* die Hochschätzung geringer Leistungsstreuungen bei der Bewertung von Bildungssystemen, führe sie doch dazu, dass be-

sondere Talente, die nicht in das Standardprofil passen, entmutigt oder gar unterdrückt würden (S. 83ff.). Weitere Konkurrenzen können bestehen zwischen fachlichen und übergreifenden Kompetenzen, zwischen kurzfristigen und langfristigen Wirkungen (S. 85 ff.).

Wechselwirkungen sind aber nicht nur zwischen verschiedenen Zielen zu beachten. Die Forschung zu *Aptitude-Treatment-Interaktionen* hat bereits in den 1970er-Jahren darauf aufmerksam gemacht, dass von derselben Methode manche Schüler(gruppen) mehr profitieren als andere (S. 90ff.). So profitierten leistungsstarke Schüler/innen in einer Studie von fachlich besonders kompetenten Lehrer/innen, während leistungsschwache bei ihnen nicht so gut lernen – sie machten größere Fortschritte bei sozial kompetenten Lehrer/innen (S. 89). Allerdings vereinfachen so globale Kategorien das Bild sehr und führen manchmal auch in Sackgassen, z. B. wenn Verlage rosa Lesebücher für Mädchen und blaue für Jungen zusammenstellen, obwohl es viele Mädchen gibt, die sich für Sachtexte interessieren, und andererseits auch Jungen, die gerne Geschichten lesen. Die Grundeinsicht aber ist wichtig: Pädagogische Maßnahmen und didaktische Materialien wirken nicht technisch einsinnig. So verschlechterten sich Schüler/innen, die sich als selbstbestimmte Personen sahen, wenn der Mathematikunterricht lehrerzentriert war; Schüler/innen mit geringem Vorwissen wiederum profitierten besonders, wenn der Unterweisung durch die Lehrperson eine Phase eigenständiger Erkundung vorgeschaltet wurde (S. 95).

Zhaos Fazit: „Pädagogische Maßnahmen … können sich positiv auf einige Schüler/innen auswirken, aber anderen schaden. … sie können die Ergebnisse in einigen Zieldimensionen verbessern und gleichzeitig andere beeinträchtigen." (S. 88)

Kontextabhängige Wirkung von Interventionen: *Bildungsgutscheine*

Dabei können auch Unterrichtsbedingungen die Wirkung einer Methode beeinflussen, z. B. abhängig davon, ob dieselben Aufgaben einzeln, in Partnerarbeit, in kleinen oder großen Gruppen zu bearbeiten sind (S. 94).

Am Beispiel von den – im Vorschulbereich auch in Deutschland umstrittenen – Bildungsgutscheinen für eine freie Schulwahl verdeutlicht *Zhao*, dass ebenso außerschulische Kontextfaktoren Einfluss haben können auf die Wirkung bildungspolitischer Maßnahmen: Werden die Schulen öffentlich finanziert, sind die Effekte auf die Schülerleistungen positiver als bei privater Finanzierung (99). Aber auch die Verfügbarkeit von Verkehrsmitteln und vor allem der Bildungshintergrund der Familien und die Finanzkraft von Eltern spielen eine Rolle dafür, ob bzw. wie eine freie Schulwahl genutzt wird (S. 99f.).

Zhaos Fazit: „Merkmale der Eltern und der Familien, in denen Kinder leben, können sich darauf auswirken, ob die freie Wahl positiv, null oder negativ wirkt." Und verallgemeinert: „Was für den einen als Heilmittel wirkt, ist für den anderen Gift." (S. 105)

Unproduktive Pendelschwünge: *New Math* und *Phonics vs Whole language*

Angesichts der vorgetragenen Befunde sind die „didaktischen Kriege" über die beste Methode des Lesen- und Schreibenlernens oder des Mathematikunterrichts unproduktiv. Bekanntlich endete der Methodenstreit zwischen Ganzheitlern und Synthetikern in Deutschland schon vor 50 Jahren in einem Patt. Aber nicht weil beide Leselehr-Methoden gleich gut waren, sondern weil beide neben ihren spezifischen Stärken auch jeweils besondere Schwächen hatten – ähnlich wie *„phonics"* und *„whole language"* in den USA (S. 106f.). Analog verlief die Auseinandersetzung zwischen „Rechnern" und „Denkern" beim Streit über die Einführung der Mengenlehre (S. 105f.). Die Suche nach einem Allheilmittel führt zu den ständig wiederkehrenden Pendelschwüngen, die in der Praxis keinen Fortschritt bringen (S. 110ff.). Statt ein gescheitertes Patentrezept gegen ein anderes auszutauschen, solle man in kleinen Schritten an Verbesserungen arbeiten. Nur ein echtes Interesse an den Problemen eines Ansatzes ermöglicht, diesen zu verbessern (S. 114).

Jede Methode hat ihre Risiken – nicht anders die Alternativen, die angeboten werden, diese Schwächen zu überwinden. Weder die Hoffnung auf Patentlösungen noch die Verständigung auf einen „goldenen Mittelweg" hilft weiter. Sie verhindern sogar Fortschritte, weil man dann nicht mehr an den spezifischen Problemen arbeitet, um sie zu überwinden, und auch nicht nach besseren Passungen für unterschiedliche Gruppen bzw. Situationen sucht (S. 114).

In der Medizin gibt es gesetzliche Vorschriften, dass Nebenwirkungen erfasst und berichtet werden müssen (S. 33). So konnte in den USA etwa der Contergan-Skandal verhindert werden, weil – anders als in Europa – das Medikament Thalidomid für Schwangere nicht zugelassen wurde. Denn nach einem ähnlichen Drama in den 1930er-Jahren war dort 1938 das *Food, Drug and Cosmetic Act* erlassen worden, das eine Offenlegung und Bewertung von Nebenwirkungen verlangt (S. 34). Anders als in der Medizin gibt es in der Pädagogik aber keine Tradition, systematisch nach Nebenwirkungen zu suchen (S. 115).

Zhaos Fazit: „Vollständigere Informationen über die Auswirkungen und Nebenwirkungen von Bildungsansätzen können die ideologischen Kriege nicht stoppen, aber sie können helfen, den ewigen Pendelschwung, das bloße Recycling alter Ideen, aufzuhalten." (S. 123)

Aufruf: Mehr Aufmerksamkeit für mögliche Nebenwirkungen!

Zhaos Botschaft, kurz zusammengefasst: Pädagogische Interventionen wirken nicht einsinnig. Ihre Wirkungen sind differenziert zu beurteilen, indem man sie auf konkurrierende Ziele bezieht, ungewollte Nebenwirkungen berücksichtigt und die Streuung der Effekte in Abhängigkeit vom Kontext bedenkt.

Er fordert deshalb eine systematischere Untersuchung von Nebenwirkungen (S. 118ff.). Davon erhofft er sich,

- die Aufgabe des illusionären Anspruchs an Programme und Methoden, sie müssten in jeder Hinsicht besser sein als ihre Konkurrenten
- eine kritischere Prüfung von Erfolgsversprechen neuer Ansätze oder Materialien durch die Nutzer/innen
- eine schrittweise Verbesserung von Programmen und Methoden statt der regelmäßigen Pendelschwünge von einem Extrem ins andere
- bei der Umsetzung in die Praxis eine größere Aufmerksamkeit für kontextbedingte Risiken der gewählten Programme, Methoden oder Materialien.

Es sei Aufgabe von Politik und Stiftungen, entsprechende Anforderungen für die Finanzierung oder Zulassung von neuen Konzepten bzw. Materialien zu formulieren. Forschung und Entwicklung müssten ihren Fokus ebenso erweitern wie Medien und Entscheider bei der Prüfung von Angeboten. Ein Appell, der nicht nur in den USA berechtigt ist, sondern auch in der deutschen Reformdebatte Beachtung verdient.

Dabei geht es im Kern um die Frage, welchen Geltungsanspruch empirische Bildungsforschung gegenüber Politik und Praxis beanspruchen kann. *Zhaos* Antwort ist klar: Forschung muss Bildungspolitik und Schulpraxis als kompetente Gegenüber respektieren, sie muss ihre Studien so anlegen und die Befunde so aufbereiten, dass diese selbstständig urteilen und unter ihren spezifischen Bedingungen erfolgreich handeln können – eine Sicht, die an einer vergessenen Tradition der 1970er-Jahre anknüpfen könnte: Evaluation als „Dienstleistung" (Brügelmann 1976) für die *„experimental colleagues"* (Stenhouse 1975) in der Praxis und eine Forschung, die nicht qua wissenschaftlicher Autorität dichotome Urteile (wirksam vs. unwirksam) fällt, sondern über die „Potenziale" (Ben-Peretz 1975) von Interventionen aufklärt und Bedingungen benennt, unter denen sie möglichst gut zur Entfaltung gebracht und die unvermeidlichen Risiken minimiert werden können.

Wie die eingangs zitierte Untersuchung zeigt, sind wir immer noch weit davon entfernt, diesen Ansprüchen gerecht zu werden. Stattdessen halten wir an der Illusion fest, didaktische Streitfragen durch Mittelwertvergleiche in Entscheidungsexperimenten lösen zu können. Die Kritik von Zhao ermuntert, den Blick zu weiten.

Anmerkung

[1] Eine Vorstellung des Buches von Yong Zhao: What works may hurt. New York: Teachers College Press 2018. Eine Kurzfassung ist erschienen. In: Die Deutsche Schule, 111. Jg., H. 2, 260-262. Die Zitate sind vom Autor (mit Unterstützung durch das Online-Programm deepL) übersetzt worden.

Literatur

Ben-Peretz, M.: The concept of curriculum potential. In: Curriculum Theory Network 5 (1975), No. 2, S. 151-159.

Brügelmann, H.: Curriculumevaluation – eine Dienstleistung für die Unterrichtspraxis. In: Seybold, H. (Hrsg.): Innovation im Unterricht. Curriculumentwicklung und handlungsorientierte Forschung. Ravensburg 1976, S. 75-93.

Brügelmann, H.: Vermessene Schulen – standardisierte Schüler. Zu Risiken und Nebenwirkungen von PISA, Hattie, VerA & Co. Weinheim/Basel 2015.

Elliott, J./Kushner, S.: The need for a manifesto for educational programme evaluation. In: Cambridge Journal of Education 37 (2007), No. 3, S. 321-336.

Hattie, J. A. C.: Visible Learning. A synthesis of over 800 meta-analyses relating to achievement. London 2009.

House, E. R.: Evaluating with validity. Beverly Hills/London 1980.

Kuhl, T./Röhr-Sendlmeier, U. M.: Der Verlauf des Rechtschreib-Lernens. Drei Didaktiken und ihre Auswirkungen auf Orthographie und Motivation in der Grundschule. Vortrag und Posterpräsentation auf dem 4. Dortmunder Symposium der Empirischen Bildungsforschung (TU Dortmund) 4.-5.7. 2018. (a)

Kuhl, T./Röhr-Sendlmeier, U. M.: Rechtschreiberfolg nach unterschiedlichen Didaktiken. Eine kombinierte Längsschnitt-Querschnittstudie in der Grundschule. Posterpräsentation auf dem Bundeskongress für Schulpsychologie (Universität Frankfurt), 21.9.2018. (b)

Sackett, D. L., et al.: Evidencebased medicine: How to practice and teach EBM. Edinburgh [1]1997, [2]2000.

Stenhouse, L.: An introduction to curriculum research and development. London 1975, dt. Zusammenfassung in: Zeitschrift für Pädagogik 19 (1975), S. 447-452.

Prof. Dr. Hans Brügelmann
Prof. em. Universität Siegen
Grundschulpädagogik und -didaktik
Engagement im Grundschulverband
Landesgruppe Bremen
hans.bruegelmann@gmx.de